Machine learning force fields based on local parametrization of dispersion interactions: Application to the phase diagram of C_{60}

Heikki Muhli,[1] Xi Chen,[1,*] Albert P. Bartók,[2] Patricia Hernández-León,[3]
Gábor Csányi,[4] Tapio Ala-Nissila,[5,6] and Miguel A. Caro[3,†]

[1]*Department of Applied Physics, Aalto University, 02150, Espoo, Finland*
[2]*Department of Physics and Warwick Centre for Predictive Modelling, School of Engineering, University of Warwick, Coventry CV4 7AL, United Kingdom*
[3]*Department of Electrical Engineering and Automation, Aalto University, 02150, Espoo, Finland*
[4]*Engineering Laboratory, University of Cambridge, Cambridge CB2 1PZ, United Kingdom*
[5]*Department of Applied Physics, QTF Center of Excellence, Aalto University, 02150, Espoo, Finland*
[6]*Interdisciplinary Centre for Mathematical Modelling and Department of Mathematical Sciences, Loughborough University, Loughborough, Leicestershire LE11 3TU, United Kingdom*

(Dated: July 16, 2021)

We present a comprehensive methodology to enable addition of van der Waals (vdW) corrections to machine learning (ML) atomistic force fields. Using a Gaussian approximation potential (GAP) [Bartók *et al.*., Phys. Rev. Lett. **104**, 136403 (2010)] as baseline, we accurately machine learn a local model of atomic polarizabilities based on Hirshfeld volume partitioning of the charge density [Tkatchenko and Scheffler, Phys. Rev. Lett. **102**, 073005 (2009)]. These environment-dependent polarizabilities are then used to parametrize a screened London-dispersion approximation to the vdW interactions. Our ML vdW model only needs to learn the charge density partitioning implicitly, by learning the reference Hirshfeld volumes from density functional theory (DFT). In practice, we can predict accurate Hirshfeld volumes from the knowledge of the local atomic environment (atomic positions) alone, making the model highly computationally efficient. For additional efficiency, our ML model of atomic polarizabilities reuses the same many-body atomic descriptors used for the underlying GAP learning of bonded interatomic interactions. We also show how the method enables straightforward computation of gradients of the observables, even when these remain challenging for the reference method (e.g., calculating gradients of the Hirshfeld volumes in DFT). Finally, we demonstrate the approach by studying the phase diagram of C_{60}, where vdW effects are important. The need for a highly accurate vdW-inclusive reactive force field is highlighted by modeling the decomposition of the C_{60} molecules taking place at high pressures and temperatures.

I. INTRODUCTION

Interatomic interactions that emanate from the underlying quantum-mechanical nature of matter can often be broken down into effective contributions in terms of classical force fields. Traditionally, these contributions are four: covalent (or "bonded"), electrostatics, Pauli repulsion, and dispersion (or "van der Waals"). In the context of classical molecular dynamics (MD) and the development of atomistic force fields, covalent interactions typically account for the portion of the potential energy surface (PES) of a system of interacting atoms that can be parametrized into a set of harmonic-like short-range functions. Electrostatic interactions, in their simplest implementation, take the form of point charges located at the positions of the atomic nuclei, interacting via a long-range Coulomb potential; more complicated forms can account for atom-centered multipole expansions and polarizabilities. Repulsion (often termed "exchange" or "Pauli" repulsion) is the very strong interaction preventing two atoms from coming too close to one another, a phenomenon that can be ascribed to Pauli's exclusion principle, and decays very rapidly with interatomic separation. Finally, "dispersion" is an umbrella term of sorts for "the rest", i.e., the difference between the quantum-mechanical energies and forces and the sum of the covalent, electrostatic and repulsion energies and forces [1].

Dispersion interactions, also often referred to as "van der Waals" (vdW) interactions (and making up part of the "correlation" energy, as referred to within the electronic structure community), are long ranged, but typically *individually* weak. E.g., the dispersion energy between two isolated atoms or molecules is usually much smaller than their corresponding electrostatic interaction. However, while electrostatic interactions may cancel out due to the balance between attractive and repulsive energy contributions, dispersion interactions are typically attractive. Thus, the dispersion interaction can be the driving force behind interesting emerging physical phenomena, such as the bonding of 2D layers of material to form 3D solids (graphite or black phosphorus being prime examples). Trending topics in nanoscience and nanotechnology include "van der Waals heterostructures" [2] and, more generally, simply the desire to describe atomistic systems to a greater level of detail.

Machine learning (ML) in molecular and materials modeling [3–5] is an emerging interdisciplinary field that touches on physics, chemistry, materials science and biochemistry, and has seen exponential growth in recent years. Still, dispersion interactions remain elusive. This is particularly true for materials modeling, where the greater atomic density complicates the description of the system. At the root of the issue is the need for local atomic descriptors as input for ML potentials to keep the atomistic problem tractable [6, 7]. In ML atomistic modeling, the representation of the whole system is carried

* xi.6.chen@aalto.fi
† mcaroba@gmail.com

out in practice by representing the individual building blocks and then adding them together [8, 9]. One of these individual blocks can be for example an atom and all of its neighbors within a given cutoff sphere of a few ångströms. This poses a serious problem for ML potentials because the effective range of vdW interactions for the system at hand can be in excess of a few *tens* of ångströms.

To circumvent this issue, a promising solution is to rely on simple analytical forms for the long-range dispersion interactions between a pair of atoms, which are computationally cheap to evaluate, and condense all the complicated many-body physics into the *local parametrization* of said function. This strategy is in the spirit of existing popular "dispersion corrections" to density functional theory (DFT) [10]. One of these approaches, the Tkatchenko-Scheffler (TS) vdW correction [11], is precisely the starting point for our present approach. We will show in this paper that TS corrections, which rely on computationally expensive integration of the charge density field, can be effectively and accurately machine learned. Furthermore, we will show how properties that are not straightforward to compute within the context of the reference method (e.g., the gradients of these charge integrals) can be readily and inexpensively obtained within our ML framework.

Previous recent attempts at machine learning vdW corrections have explored the ideas of mixing parametric and non-parametric fits [12], learning highly accurate dispersion energies to add them directly on top of DFT [13], or applying preexisting (unmodified) vdW correction schemes on top of ML potentials [14]. The most similar in spirit to our current approach is the work by Bereau *et al.* [12], who developed accurate "all-ML" interatomic potentials for small systems, and used a local parametrization of a physical vdW model (many-body dispersion [15]), which goes beyond the TS model used here. Some of these models achieve highly accurate results but still fall short of "force field" computational efficiency, and are therefore not amenable to large-scale MD simulations. Therefore, the focus in this paper is on adding vdW corrections within a general and flexible framework, which can be easily extended in the future to accommodate more sophisticated vdW models and other interactions (such as electrostatics), while maintaining a very low computational footprint. This allows us to perform large-scale MD simulations, both with regards to the number of atoms and accessible time scales, albeit necessarily sacrificing some accuracy in the process.

The outline of this paper is as follows. We first introduce the general methodology for the addition of pairwise dispersion corrections on top of DFT energies. This is followed by a general discussion of the ML methods we use to construct force fields, namely GAPs, kernels and many-body atomic descriptors. Then we discuss how this methodology is applied to learning the dispersion energies from local parameters of the atomic environments and how we calculate the dispersion forces analytically. Before moving on to the simulation results, we also discuss the implementation of our methodology in the TurboGAP and QUIP codes. We then present a new GAP force field for carbon with vdW corrections, specifically tailored for simulation of C_{60} based systems, and show basic performance and accuracy tests. Finally, with our new GAP, we characterize the phase stability and phase transformations taking place in the C_{60} molecular system over a wide range of pressures and temperatures.

II. METHODOLOGY

A. Tkatchenko-Scheffler vdW method

The TS method relies on two fundamental approximations, one regarding the functional form of dispersion interactions and another regarding the parametrization of said functional form [11, 16, 17]. First, there is a range separation of the electron correlation energy, where the short-range correlation is captured by the underlying DFT functional (usually, but not necessarily, PBE [18]) and the long-range correlation is modeled via the London dispersion formula for pairwise interactions with polynomial decay as $\propto 1/r^6$ [16]. The transition between the short and long ranges is modeled via a "damping" function, an approach first introduced by Grimme [10]. Second, the novelty in TS resides in how the London dispersion formula and the damping function are parametrized, taking the atomic environments of the two interacting atoms into account. TS assumes an "atom in a molecule" approach, which approximates the properties of the atoms in a molecule or solid as proportional to the properties of the "free" (neutral) atom. The resulting equations are remarkably simple:

$$
\begin{aligned}
E_{\mathrm{TS}} &= -\frac{1}{2}\sum_i \sum_{j\neq i} C_{6,ij}\frac{f_{\mathrm{damp}}(r_{ij}; r_{ij}^{\mathrm{vdW}}, s_{\mathrm{R}}, d)}{r_{ij}^6},\\
C_{6,ij} &= \frac{2C_{6,ii}C_{6,jj}}{\frac{\alpha_{j,0}}{\alpha_{i,0}}C_{6,ii} + \frac{\alpha_{i,0}}{\alpha_{j,0}}C_{6,jj}}, \quad C_{6,ii} = {\nu_i}^2 C_{6,ii}^{\mathrm{free}},\\
\alpha_{i,0} &= \nu_i \alpha_{i,0}^{\mathrm{free}}, \quad r_{ij}^{\mathrm{vdW}} = r_i^{\mathrm{vdW}} + r_j^{\mathrm{vdW}}, \quad r_i^{\mathrm{vdW}} = {\nu_i}^{\frac{1}{3}} r_{i,\mathrm{free}}^{\mathrm{vdW}},\\
\nu_i &= \frac{V_i}{V_i^{\mathrm{free}}} = \frac{\int \mathrm{d}\mathbf{r}\, w_i(\mathbf{r})\rho(\mathbf{r})}{\int \mathrm{d}\mathbf{r}\, \rho_i^{\mathrm{free}}(\mathbf{r})}, \qquad w_i(\mathbf{r}) = \frac{\rho_i^{\mathrm{free}}(\mathbf{r})}{\sum_j \rho_j^{\mathrm{free}}(\mathbf{r})},\\
f_{\mathrm{damp}}(r_{ij}; r_{ij}^{\mathrm{vdW}}) &= \frac{1}{1+\exp\left[-d\left(\frac{r_{ij}}{s_{\mathrm{R}} r_{ij}^{\mathrm{vdW}}} - 1\right)\right]}.
\end{aligned}
\tag{1}
$$

Here, the local environment dependence of the force constants $C_{6,ij}$, van der Waals radii r_{ij}^{vdW} and atomic polarizabilities α_0 is taken into account via the (effective) Hirshfeld volumes ν_i, given as a functional of the electron density (computed at the DFT level). The various parameters labeled as "free" refer to the reference quantities of the free (isolated and neutral) atoms, and are tabulated [10, 19]. The damping function used to screen the TS force field also relies on two empirical parameters, $d = 20$ and $s_{\mathrm{R}} = 0.94$ (for PBE; s_{R} is functional dependent [11, 20, 21]). All other quantities within the TS framework are extracted directly from the electron density via the Hirshfeld volumes.

For actual simulations to be manageable, a vdW cutoff radius needs to be introduced at a distance where vdW interactions are expected to be negligible, to make the sums over

atom pairs finite. The cost of computing pairwise vdW interactions grows approximately as the cube of this cutoff radius. Therefore, in addition to the TS damping function that screens the dispersion interaction at short distances, we also introduce a smoothing function that makes the dispersion energies and their derivatives, that is, the dispersion forces, smooth and continuous at the vdW cutoff radius, and also takes care of the singularity at $r_{ij} \to 0$. For computational reasons, we employ a polynomial function for this purpose:

$$E_{\mathrm{TS}} = -\frac{1}{2}\sum_i \sum_{j\neq i} C_{6,ij}\frac{f_{\mathrm{damp}}(r_{ij})}{r_{ij}^6} f_{\mathrm{cut}}(r_{ij}; r_{\mathrm{cut}}, d_{\mathrm{b}}),$$

$$f_{\mathrm{cut}}(r_{ij}) = \begin{cases} 1, & \text{if } r_{\mathrm{c,i}} + d_{\mathrm{b,i}} < r_{ij} \leq r_{\mathrm{c,o}} - d_{\mathrm{b,o}} \\ 1 - 3r_{\mathrm{b,o}}^2 + 2r_{\mathrm{b,o}}^3, & \text{if } r_{\mathrm{c,o}} - d_{\mathrm{b,o}} < r_{ij} \leq r_{\mathrm{c,o}} \\ 3r_{\mathrm{b,i}}^3 - 2r_{\mathrm{b,i}}^3, & \text{if } r_{\mathrm{c,i}} < r_{ij} \leq r_{\mathrm{c,i}} + d_{\mathrm{b,i}} \\ 0, & \text{otherwise,} \end{cases}$$

$$r_{\mathrm{b,o}} = \frac{r_{ij} - r_{\mathrm{c,o}} + d_{\mathrm{b,o}}}{d_{\mathrm{b,o}}}, \quad r_{\mathrm{b,i}} = \frac{r_{ij} - r_{\mathrm{c,i}}}{d_{\mathrm{b,i}}}, \tag{2}$$

where $r_{\mathrm{c,o}}$ is the cutoff radius for the vdW interaction and $d_{\mathrm{b,o}}$ is the width of the "buffer" region where the cutoff function smoothly switches from 1 to 0 as it approaches $r_{\mathrm{c,o}}$. The same approach, in this case switching from 0 to 1, is used for an inner cutoff $r_{\mathrm{c,i}}$ at small interatomic separations (usually with an inner buffer region $d_{\mathrm{b,i}}$ between 0.5 and 1 Å), to avoid the singularity in the original TS expression for the damping function as $r_{ij} \to 0$. The combined effect of the damping function and cutoff functions can be visualized in the Supplemental Information (SI).

A clear limitation of the TS method is that it neglects many-body effects [15], which can be important in particular for accurately estimating the C_6 coefficients ("electronic" many-body effects) [22]. The D3 method [23], which also omits many-body effects, can yield more accurate dispersion energies than TS in some cases and would be similarly computationally cheap. Our choice of TS as dispersion correction scheme for this work serves two purposes. On the one hand, it demonstrates that vdW corrections can be efficiently coupled to ML potentials and used in large-scale MD simulation. On the other, it paves the way for subsequent improvements that directly feed on effective Hirshfeld volumes (such as MBD) or can benefit from local parametrization of a long-range interaction with simple functional form, the most relevant of which would be long-range electrostatics.

B. Gaussian approximation potentials

Within the context of the Gaussian approximation potential (GAP) framework [9, 24], which is an ML approach for atomistic modeling based on kernel ridge regression (KRR), any physical property f that can be assigned to a local atomic environment, such as a local energy, a force or an effective Hirshfeld volume, can be written as a linear combination of kernel functions centered on the training points:

$$\bar{f}_*(\mathbf{d}_*) = \delta^2 \sum_t \alpha_t k(\mathbf{d}_*, \mathbf{d}_t), \tag{3}$$

where $\mathbf{d}_*$ is the local atomic descriptor for which the prediction is made, $\mathbf{d}_t$ are the descriptors of the configurations in the training set, $k(\mathbf{d}_*, \mathbf{d}_t)$ is a kernel, or similarity measure (bounded between 0 and 1) between the atomic environments characterized by $\mathbf{d}_*$ and $\mathbf{d}_t$, α_t are a set of fitting coefficients, and δ is a scaling parameter with units of the magnitude under study (e.g., for energies δ is given in eV). Due to the large amount of available training data points, it is usually beneficial to sparsify the model such that many fewer fitting coefficients α_t in Eq. (3) are used for prediction than there are local environments in the training database (since the cost of evaluating the model scales linearly in the size of the sparse set) [24]. The atomic environments present in the training database are also often repetitious, which means that they do not contribute much new information to the model compared to the computational cost of including them.

More generally, a GAP can be constructed by addition of several terms of the form of Eq. (3), typically (but not limited to), two-body (2b), three-body (3b) and many body (mb). A more detailed account of GAP construction, including discussion of descriptors, sparsification, regularization and training, has been given elsewhere [6, 9, 24–26]. For conciseness, here we restrict ourselves to giving a brief overview of the mb descriptors used for learning effective Hirshfeld volumes in the context of the present methodology, and refer the reader to the literature for further details.

The numerical fingerprint of an atomic environment centered on a particular atom can be obtained from the smooth overlap of atomic positions (SOAP) [6]. For the SOAP mb descriptor, we define the kernel directly as proportional to the overlap integral of two atomic density fields $\rho(\mathbf{r})$ and $\rho'(\mathbf{r})$ [6, 26] averaged over all possible relative orientations $\hat{R}$ of the two atomic environments:

$$k(\rho, \rho') = \int \mathrm{d}\hat{R} \left| \int \mathrm{d}\mathbf{r}\, \rho(\mathbf{r}) \rho'(\hat{R}\mathbf{r}) \right|^n, \tag{4}$$

where the atomic density field is constructed from the sum of individual atomic contributions within the atomic neighborhood (as defined by a cutoff distance) of the central atom i, possibly including the central atom itself:

$$\rho(\mathbf{r} - \mathbf{r}_i; i) = \sum_j g(\mathbf{r} - \mathbf{r}_i; \mathbf{r}_{ij}, \{\lambda\}) \qquad \forall \quad r_{ij} < r_{\mathrm{SOAP}}, \tag{5}$$

where g is a smearing function (e.g., a Gaussian) centered at the position of each atomic neighbor [6, 26]. We have collectively represented the different SOAP hyperparameters, which determine the fine detail of the atomic representation, by $\{\lambda\}$. The integral over relative orientations in Eq. (4) ensures the rotational invariance of this kernel, which also satisfies the permutational and translational invariance criteria. The integrand is raised to power $n \geq 2$ to retain all the angular information of the original environments [6]. To make the integral computationally tractable, the atomic density field is expanded in

terms of a radial basis $\{g_n\}$ and spherical harmonics Y_{lm}:

$$\rho(\mathbf{r}) \equiv \sum_{nlm} c_{nlm}\, g_n(r) Y_{lm}(\theta, \phi), \tag{6}$$

with the normalized power spectrum of the expansion coefficients c_{nlm} giving the SOAP vectors:

$$p_{nn'l} = \sum_m c_{nlm} (c_{n'lm})^*, \qquad \mathbf{q} = \frac{\mathbf{p}}{\sqrt{\mathbf{p} \cdot \mathbf{p}}}. \tag{7}$$

The final (normalized) form of the SOAP kernel between atomic environments i and j is given by a dot product [6, 26]:

$$k^{\mathrm{SOAP}}(i, j) \equiv \mathbf{q}_i \cdot \mathbf{q}_j, \tag{8}$$

where i refers to the atom with density ρ and j to the atom with density ρ' in Eq. (4). In this work we use the formulation of SOAP introduced in Ref. [26], which offers several computational advantages. SOAP is particularly well suited for the representation of atom-based properties that depend on the entire (many-body) local atomic environment.

C. ML model of Hirshfeld volumes and gradients

In our vdW model, SOAP descriptors are used to predict the effective Hirshfeld volumes of the different atoms based on their local atomic environments. These volumes can then be used to calculate the TS dispersion correction using Eq. (1). The effective Hirshfeld volume for atom i is thus predicted by

$$\nu_i = \sum_{s \in S} \alpha_s |k^{\mathrm{SOAP}}(i, s)|^\zeta, \tag{9}$$

where α_s are the fitting coefficients obtained from the ML model and S denotes the chosen sparse set of atomic environments, which is a subset of the whole training set T. We implicitly assume $\delta = 1$ [cf., Eq. (3)]. The parameter ζ is empirical and takes small positive values $\zeta \geq 1$. It is used to make the kernels sharper to effectively emphasize the differences between environments and to accentuate the sensitivity of the kernel to changing atomic positions [6]. To further improve the fitting, the reference DFT data can be shifted by approximately the mean of the whole database:

$$\nu_i = \nu_i^{\mathrm{DFT}} - \nu_0, \quad \nu_0 = \frac{1}{N} \sum_{i=1}^{N} \nu_i^{\mathrm{DFT}}. \tag{10}$$

Subtraction of ν_0 is done for the training stage (so that the quantity to be learned is smoother) and then added back at the prediction stage. With these predicted Hirshfeld volumes we can calculate the pairwise dispersion energies using the TS dispersion correction given by Eq. (1) and Eq. (2).

The dispersion forces are given as the negative gradients of the dispersion energy:

$$f_k^\alpha = -\frac{\partial E_{\mathrm{TS}}}{\partial r_k^\alpha}, \tag{11}$$

where k denotes the atom and α denotes the Cartesian coordinate. Here we give the result of the differentiation (see SI for derivation details):

$$\begin{aligned} f_k^\alpha = \sum_i \frac{\partial \nu_i}{\partial r_k^\alpha} \sum_{j \neq i} \Bigg[& \frac{C_{6,ij}}{\nu_i} \frac{f_{\mathrm{damp}}(r_{ij})}{r_{ij}^6} f_{\mathrm{cut}}(r_{ij}) \\ & - \frac{r_{ij}}{(r_{ij}^{\mathrm{vdW}})^2} \frac{d}{s_R} f_{\mathrm{damp}}(r_{ij})^2 f_{\mathrm{cut}}(r_{ij}) \\ & \times \exp\left(-d\left(\frac{r_{ij}}{s_R r_{ij}^{\mathrm{vdW}}} - 1\right)\right) \frac{C_{6,ij}}{r_{ij}^6} \frac{1}{3\nu_i^{2/3}} r_{i,\mathrm{free}}^{\mathrm{vdW}} \Bigg] \\ & + \sum_i \delta_{ik} \sum_{j \neq i} \frac{C_{6,ij}}{r_{ij}^7} f_{\mathrm{damp}}(r_{ij}) f_{\mathrm{cut}}(r_{ij}) (r_j^\alpha - r_i^\alpha) \\ & \times \left[\frac{6}{r_{ij}} - \frac{d}{s_R r_{ij}^{\mathrm{vdW}}} f_{\mathrm{damp}}(r_{ij}) \exp\left(-d\left(\frac{r_{ij}}{s_R r_{ij}^{\mathrm{vdW}}} - 1\right)\right) \right] \\ & + \sum_i \delta_{ik} \sum_{j \neq i} \frac{C_{6,ij}}{r_{ij}^7} f_{\mathrm{damp}}(r_{ij}) (r_j^\alpha - r_i^\alpha) D_{ij}. \end{aligned} \tag{12}$$

where δ_{ik} is the Kronecker delta and D_{ij} is defined below. We note that the expression above includes terms whose differentiation is trivial, but also terms which depend on the gradient of the Hirshfeld volume, which is a functional of the electron density. Since this functional is non-variational with respect to changes in the charge density (i.e., the ν_i are not obtained via minimization, but simply by numerical integration), the Hellmann-Feynman theorem does not apply. Consequently, most DFT codes do not report the contribution of the gradients of the Hirshfeld volumes to the dispersion forces.

By contrast, in our method the gradient of the Hirshfeld volume with respect to the position of atom k is straightforward to compute in terms of kernel and descriptor derivatives:

$$\frac{\partial \nu_i}{\partial r_k^\alpha} = \sum_{s \in S} \alpha_s \zeta |k(i, s)|^{\zeta - 1} \mathbf{q}_s \cdot \frac{\partial \mathbf{q}_i}{\partial r_k^\alpha}. \tag{13}$$

The derivatives of the SOAP descriptors are readily available from any code used for regular cohesive energy GAP computations which is able to compute forces. In our case, these derivatives are implemented in the GAP [24] and TurboGAP [26] codes. Finally, the coefficient D_{ij} that appears due to the derivative of the smoothing function $f_{\mathrm{cut}}(r_{ij})$ on the final line of Eq. (12) is given by:

$$D_{ij} = \begin{cases} -\frac{6}{d_{\mathrm{b,o}}}(-r_{\mathrm{b,o}} + r_{\mathrm{b,o}}^2), & \text{if } r_{\mathrm{c,o}} - d_{\mathrm{b,o}} < r_{ij} \leq r_{\mathrm{c,o}} \\ -\frac{6}{d_{\mathrm{b,i}}}(r_{\mathrm{b,i}} - r_{\mathrm{b,i}}^2), & \text{if } r_{\mathrm{c,i}} < r_{ij} \leq r_{\mathrm{c,i}} + d_{\mathrm{b,i}} \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

with $r_{\mathrm{b,o}}$ and $r_{\mathrm{b,i}}$ given in Eq. (2).

D. TurboGAP and QUIP/GAP implementations

We have implemented the present methodology in the GAP/QUIP [27] and TurboGAP [28] codes. To predict the

dispersion energies and forces, the implementation uses the fitting coefficients $\boldsymbol{\alpha}$ and the matrix of SOAP descriptors for the sparse set $\mathbf{Q}_S$. Those are both precomputed at the training stage from the DFT reference data of Hirshfeld volumes for different atomic environments and saved to a file, which is read and stored in memory at the beginning of a new simulation. Given a new atomic structure (a molecule or a solid), the Hirshfeld volumes are predicted for each atom in the structure using their SOAP many-body descriptors, which are computed on the fly, and the precomputed $\boldsymbol{\alpha}$ and $\mathbf{Q}_S$. The predicted Hirshfeld volumes are given by Eq. (9), which is used in matrix form in TurboGAP:

$$\nu_* = (\mathbf{q}_* \mathbf{Q}_S^T)^{\odot\zeta} \boldsymbol{\alpha} + \nu_0. \tag{15}$$

Here we denote the effective Hirshfeld volume and the SOAP descriptor of a new atomic environment with the subscript "$*$". The Hadamard operator $\odot$ indicates element-wise exponentiation, in this case. If the forces are to be calculated, the derivatives of the Hirshfeld volumes are calculated in the same part of the code, reusing the different array quantities:

$$\frac{\partial \nu_*}{\partial r_k^\alpha} = \left(\zeta (\mathbf{q}_* \mathbf{Q}_S^T)^{\odot(\zeta-1)} \odot \boldsymbol{\alpha}\right) \mathbf{Q}_S \frac{\partial \mathbf{q}_*^T}{\partial r_k^\alpha}, \tag{16}$$

where $\mathbf{q}_* \in \mathbb{R}^{1\times N_{\mathrm{SOAP}}}$, $\boldsymbol{\alpha} \in \mathbb{R}^{1\times N_S}$ and $\mathbf{Q}_S \in \mathbb{R}^{N_S\times N_{\mathrm{SOAP}}}$. Here, the second Hadamard operator indicates element-wise multiplication.

Since the SOAP many-body descriptors only see the local environment within the SOAP cutoff sphere of radius r_{SOAP}, the descriptor of atom i also only has non-zero SOAP gradients for atoms within that cutoff. The SOAP descriptor of atom i gives the gradients for a SOAP neighbor (within the SOAP cutoff) atom k as $\partial \mathbf{q}_i / \partial r_k^\alpha$. The cutoff for pairwise vdW interactions r_{cut} is generally different and usually larger than r_{SOAP}, because the vdW interaction is long ranged and SOAP descriptors encode the local atomic environment only. It is technically possible, however, to define them such that r_{SOAP} is larger. Whichever cutoff radius is the larger one will determine the neighbor list that needs to be built. If atom i in the structure has n_i neighbors within r_{cut} (counting also the atom itself), and assuming it is the largest cutoff, the neighbor list will have $N_{\mathrm{pairs}} = \sum_i n_i$ elements. The following quantities are then calculated for all pairs of atoms i and j within the vdW cutoff radius:

$$C_{6,ij}, \quad f_{\mathrm{damp}}(r_{ij}; r_{ij}^{\mathrm{vdW}}, s_R, d), \quad r_{ij}^{-6}, \quad r_{ij}^{\mathrm{vdW}}, \tag{17}$$

$$\exp\left(-d\left(\frac{r_{ij}}{s_R r_{ij}^{\mathrm{vdW}}} - 1\right)\right), \quad r_j^\alpha - r_i^\alpha \tag{18}$$

and stored in vectors of size N_{pairs}. By looping over the atoms i and their neighbors j, these quantities can be used to calculate the inner sums in Eq. (12), such that we get an equation that reads

$$f_k^\alpha = \sum_i \frac{\partial \nu_i}{\partial r_k^\alpha} A_i + \sum_i \delta_{ik} B_i. \tag{19}$$

The explicit expressions for A_i and B_i can be retrieved by collecting the corresponding terms in Eq. (12). Then, by looping over atoms i and their SOAP neighbors k again, we get the final dispersion forces for the atoms. Here, the gradients of the effective Hirshfeld volumes are zero if $r_{ik} > r_{\mathrm{SOAP}}$, and those terms can be skipped in the sum. The first term involves the effect of the changing electronic density within the local environment on the atoms in that environment. The second term is the long-range part that also appears when neglecting the gradients of the Hirshfeld volumes. As we have discussed earlier, this second term gives the same TS forces that are usually reported by DFT codes, which often lack the interactions given by the first term altogether. We provide some illustrative examples of this issue and how the present implementation overcomes it in Sec. III C.

A reference implementation is provided in the QUIP package [27], which together with the GAP plugin also facilitates fitting the model of Hirshfeld volumes from a database of atomic configurations. For a database containing N atoms, with effective Hirshfeld volumes obtained from DFT calculations (shifted by the approximate mean as in Eq. (10)) $\boldsymbol{\nu}$, the set $\{\mathbf{q}_i\}_{i=1}^N$ is first computed, of which N_S representative points are selected as the sparse set. Elements of covariance matrices $\mathbf{K}_{TS}$ and $\mathbf{K}_{SS}$ are computed using the kernel definition in Eq. (3), where the subscripts T and S indicate the full training set and sparse set of atomic environments, respectively. In matrix form these are given by:

$$\mathbf{K}_{SS} = \left(\mathbf{Q}_S \mathbf{Q}_S^T\right)^{\odot\zeta}, \quad \mathbf{K}_{TS} = \left(\mathbf{Q}_T \mathbf{Q}_S^T\right)^{\odot\zeta}, \quad \mathbf{K}_{ST} = \mathbf{K}_{TS}^T, \tag{20}$$

where $\mathbf{Q}_T \in \mathbb{R}^{N_T\times N_{\mathrm{SOAP}}}$ and $\mathbf{Q}_S \in \mathbb{R}^{N_S\times N_{\mathrm{SOAP}}}$ have the SOAP descriptors as their rows. The coefficients are obtained using the sparse Gaussian process regression [29] as

$$\boldsymbol{\alpha} = (\mathbf{K}_{SS} + \mathbf{K}_{ST}\boldsymbol{\Sigma}^{-1}\mathbf{K}_{TS})^{-1}\mathbf{K}_{ST}\boldsymbol{\Sigma}^{-1}\boldsymbol{\nu}, \tag{21}$$

where $\boldsymbol{\Sigma}$ is a diagonal matrix containing the regularization parameters associated with each effective Hirshfeld volume observation. To predict the vdW interaction terms in QUIP, we first obtain the atomic force constants $C_{6,ij}$ and the vdW radii r_{ij}^{vdW}, as well as their derivatives, if required, with respect to the Cartesian coordinates of neighboring atoms. For each atomic pair within the vdW cutoff, these terms are substituted in the TS expression of the dispersion energy, while force and stress contributions are also accumulated.

The QUIP implementation, while less efficient than the alternative TurboGAP, offers the possibility of fitting and using hierarchical (Δ-learning) models, as well as Python bindings via the quippy extension of QUIP and ASE [30] integration.

For computational efficiency, the TurboGAP code offers the capability of reusing the same SOAP descriptors that are used for constructing a cohesive energy GAP for the Hirshfeld volume prediction. In practical terms, this means that the TS corrections as implemented here can be applied on top a cohesive energy force field with only a small increment in computational cost, of the order of 20-50% extra CPU time for the systems studied here with the C60 GAP. We note, however, that the computational overhead depends on many factors, such

as the effective range of vdW interactions for the system under study, the ratio of vdW cutoff to SOAP cutoff (since the scaling laws are different), the typical number of neighbors within the cutoff sphere, or the relative cost of the kernel regression step compared to building the descriptors.

III. SIMULATIONS OF CARBON-BASED SYSTEMS

As a proof of principle for the new ML implementation of vdW interactions, we study the phase diagram of the C_{60} molecular system. The interaction between individual C_{60} molecules at low pressures is almost exclusively driven by vdW interactions, whereas, as temperature and pressure increase, the role of repulsion and covalent interactions (in the form of broken and created chemical bonds) becomes more prominent. At very high pressure there is a transition from C_{60} to amorphous carbon, whereas at high temperature (but lower pressure) the C_{60} molecules decompose to graphite. Both transitions require a reactive force field to be characterized. Thus, this is a problem that requires accurate description of both vdW interactions and chemical reactions via a vdW-inclusive reactive force field. This combination of features can currently only be delivered by *ab initio* methods or the present generation of ML potentials. At the time and length scales necessary to carry out a comprehensive characterization of the phase diagram of C_{60}, *ab initio* methods are prohibitively expensive, leaving ML force fields as the only viable option.

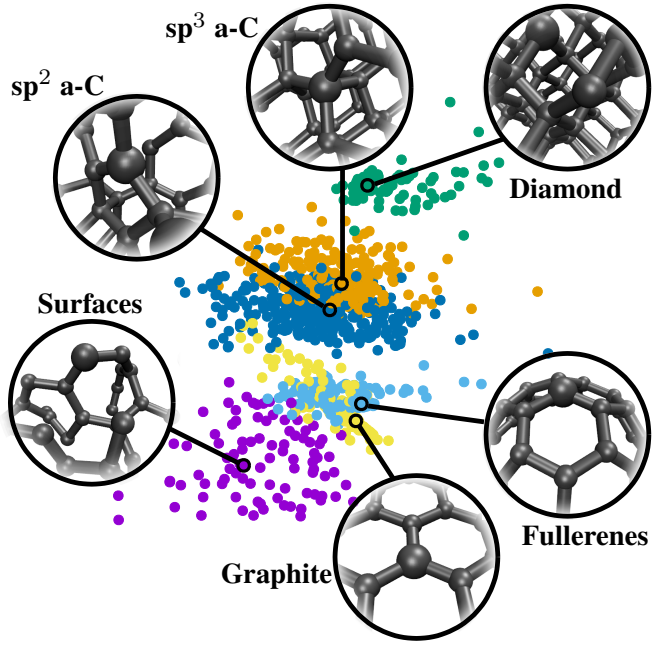

FIG. 1. Overview of the atomic configurations present in the database, using a sparse set of 1000 structures. Their composition is captured in this 2D embedding, where the distances between points mimic the SOAP dissimilarities between the corresponding atoms. We used a soap_turbo kernel with 4 Å cutoff for the dissimilarity measure and a simple hierarchy of six clusters. Several representative structures are shown for reference.

A. C60 GAP force field

To accurately describe the strong interatomic interactions in carbon materials, i.e., repulsion and covalent interactions, we constructed a new GAP force field for C, using the structural database of the 2017 amorphous carbon (a-C) GAP of Deringer and Csányi (GAP17) [25] as a baseline. GAP17 contains different amorphous and liquid C structures, dimer configurations, diamond, graphite and reconstructed surfaces. It is thus a very solid general-purpose carbon potential that gives reasonable results for most test carbon structures. However, it lacks fine resolution for some applications. For the present study, we enhanced the structural database of GAP17 with graphitic structures (exfoliation curves of graphite and bilayer graphene, and glassy carbon), some reconstructed crystalline (diamond) surfaces, and plenty of C_{60} based structures, such as distorted C_{60} monomers, undistorted C_{60} dimers and colliding C_{60} dimers. The latter are useful to correctly describe the decomposition of C_{60} molecules at high pressure. We also included some C trimer structures to improve the stability of the 3b GAP terms under high compression. For the Hirshfeld volume fit, the GAP17 database and the new C_{60} structures were used. The GAP17 and new database were computed at the DFT level of theory with the PBE [18] functional with VASP [31, 32]. The reference TS calculations (including the Hirshfeld volume computations) follow the implementation in VASP by Bučko *et al.* [33]. The composition of the database of atomic structures can be visualized in Fig. 1, where we use a new method to embed high-dimensional data in two dimensions, based on a hierarchical combination of cluster-based data classification and multidimensional scaling [34, 35]. This is a popular tool for visualizing structural databases in the context of ML applied to the study of atomic systems [36–39].

The new C60 GAP potential uses 2b descriptors with 4.5 Å cutoff, 3b descriptors with 2.8 Å cutoff and soap_turbo mb de-

TABLE I. Timings for the C60 GAP, computed on a test system made up of 64 C_{60} units at 10 kbar; see Sec. III D for a discussion on the structure. The timings exclude the time spent on neighbors lists, reading input files, etc., they only report the evaluation of energy and forces. The timings are obtained running TurboGAP on a compute node equipped with two Intel Xeon Gold 6230 CPUs at 2.1 GHz clock speed (20 physical cores per CPU). We note that these are CPU times (real time times number of cores), that is the reason why the timings increase with more cores. An example simulation of 256 C_{60} molecules (15360 atoms) over 100 ps of MD with 1 fs time step, run on 40 CPU cores with the vdW cutoff radius set to 20 Å, would take 1280 CPUh, or 32h of real time.

# cores	CPU time (ms/atom/MD step)			
	No vdW	GAP+Hirshfeld	vdW (10 Å)	vdW (20 Å)
1	1.29	1.49	1.58	2.25
5	1.48	1.57	1.64	2.23
10	1.62	1.70	1.79	2.42
20	2.11	2.18	2.31	3.07
40	2.13	2.24	2.65	3.00

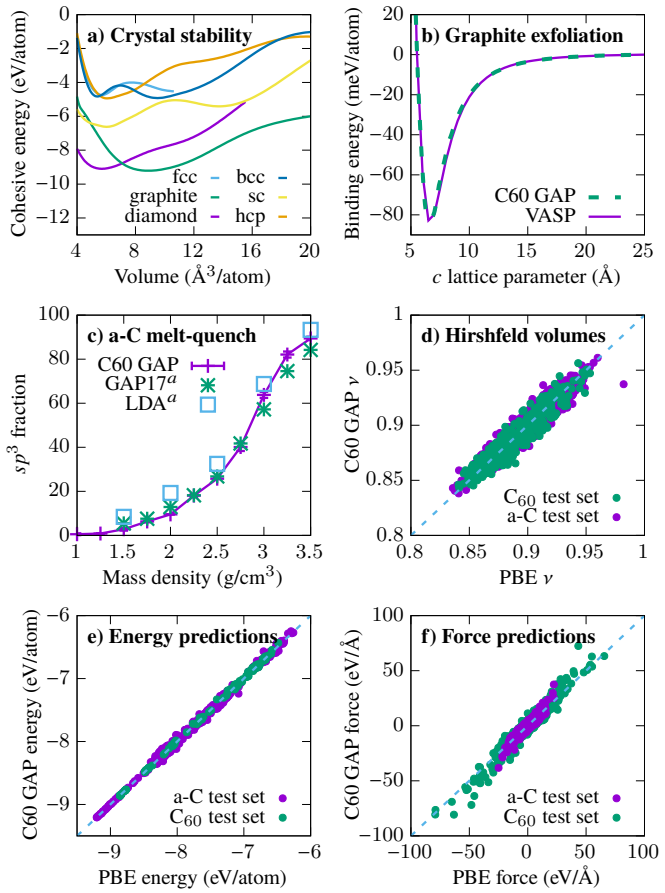

FIG. 2. Basic tests of the C60 GAP. a) Cohesive energy predictions for six crystal structures vs. volume, showing the expected stability of diamond at high pressure and graphite at low pressure. All other phases are much higher in energy. b) Graphite exfoliation curve, compared to the VASP reference data, including TS corrections with a 50 Å cutoff (the standard cutoff used by VASP). The graphite exfoliation curve shows sub-meV/atom agreement between the DFT reference data and the C60 GAP. c) a-C generation throughout different mass densities (ten 216-atom simulations per density value were used), closely following the strategy presented in Ref. [25] and comparing to data reported therein. d) Predicted Hirshfeld volumes for a-C and C_{60} test sets not included in the training set. The RMSEs are 0.0051 and 0.0046 for the a-C and C_{60} test sets, respectively. e) Energies predicted by the C60 GAP (without vdW corrections) on the a-C and C_{60} test sets vs the PBE predictions. We do not compare against the PBE+TS energies since the agreement between TS and our present methodology is fully captured by the Hirshfeld volume comparison in d); further comparison is provided in Sec. III C. The RMSEs are 37 and 21 eV/atom for a-C and C_{60} test sets, respectively. f) The same comparison as in e) but for forces in this case, with RMSEs of 1.17 and 1.69 eV/ Å for a-C and C_{60}, respectively. ^{a}Data from Ref. [25].

scriptors, as available in the GAP code, with 4.5 Å cutoff. The potential is freely available from the Zenodo repository [40] and can be used to run simulations with QUIP/GAP and TurboGAP. Besides the new vdW capabilities, this potential also incorporates a "core potential" term that ensures accurate simulation also at very close interatomic separations. E.g., the potential should remain stable for simulations of atomic collisions up to 1-2 keV, although quantitative accuracy at these energies has only been specifically enforced for the C dimer. Therefore, even though we introduce the potential as tailored for C_{60} simulation, it should also be regarded as a high-quality general-purpose potential to model carbon materials, graphitic carbons in particular. Basic tests of this potential are summarized in Fig. 2, showing very good performance across various applications. We emphasize that the errors for the a-C tests are actually smaller than those reported for GAP17 [25], on which the C60 GAP is based. This further highlights C60 GAP as an excellent general-purpose carbon potential, beyond being specifically designed to accurately model C_{60} systems.

Table I shows timings for the present C60 GAP potential. As can be seen, the potential runs rather fast for the tested system and addition of Hirshfeld volume prediction has a very small overhead of circa 5%. Adding vdW corrections results in an additional CPU cost that depends strongly on the vdW cutoff. For this test system, the overall overhead of adding vdW corrections, including Hirshfeld volume prediction, is circa 20% with a 10 Å vdW cutoff. This number grows quickly to circa 50% with a 20 Å cutoff. These two values will delimit the range of practically relevant cutoffs for most systems. In particular for carbon materials, a vdW cutoff of $\sim$ 15 Å can be considered sufficient, since the pairwise vdW interactions decay very rapidly beyond this point [see Fig. 2 b) for the case of graphite and the SI for a thorough numerical analysis in C_{60} and a-C]. We find that the TurboGAP code scales well with the number of CPU cores, and that the addition of vdW corrections does not seem to affect the scaling behavior. For the test shown in Table I, the rule of thumb is that as the number of CPU cores increases by a factor of 2, the calculation runs about a factor of x1.8 faster. In general, how far the calculation scales will depend on the number of atoms being simulated. For small systems, scaling is limited to a couple of compute nodes. For very large systems, more CPU cores can be used. We managed to get good scaling for a 1M-atom graphitic carbon system up to 2048 MPI processes (on as many physical CPU cores). Efforts on improving the software and computational performance are ongoing, and will be reported elsewhere.

B. Hirshfeld volumes: locality and learning rates

The model involves several hyperparameters that can improve the predictions but also have a huge impact on the required computational time. Two of these parameters are the SOAP descriptor cutoff radius and the size of the sparse set. The former gives the locality of the physical parameters that the model is predicting and the latter determines the learning rate of the model. Fig. 3 shows the RMSE of the predicted effective Hirshfeld volumes for amorphous carbon and C_{60} test sets as a function of SOAP cutoff radius and the sparse set size. In Fig. 3 a) and b), for a-C and C_{60} test sets respectively, it can be seen that the errors decrease quickly as the size of the training set or the sparse set is increased, but the model accuracy saturates such that increasing the sparse set size is not beneficial after a certain point due to the increased computational cost. The locality test in Fig. 3 c) shows that for the a-C test

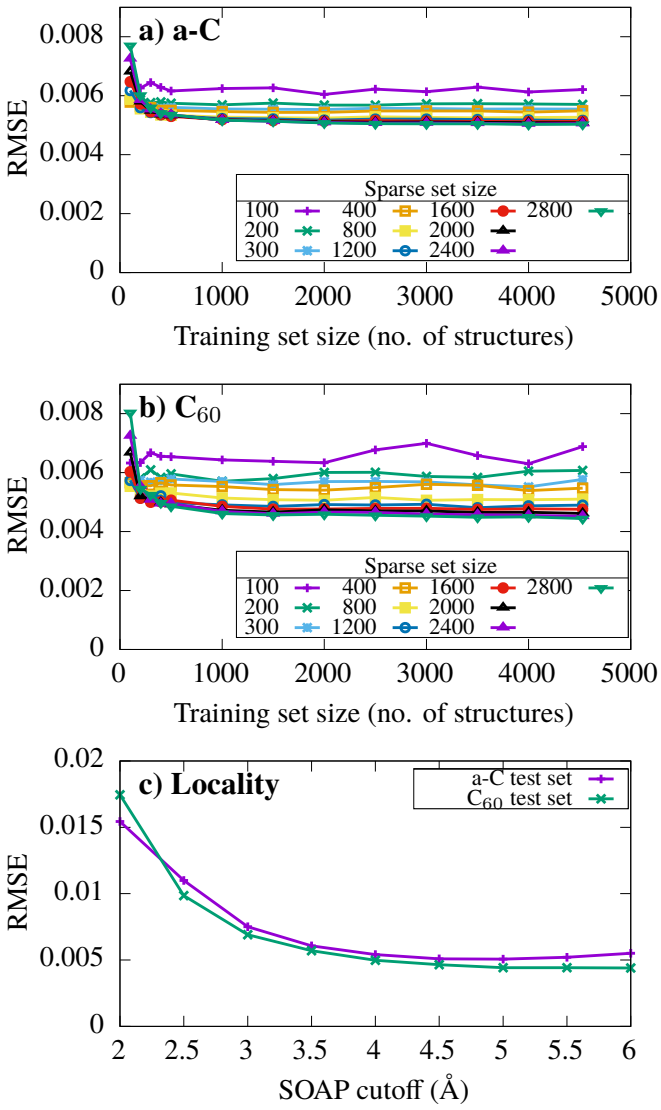

FIG. 3. Hirshfeld volume (a-b) learning rates and c) locality tests for amorphous carbon (a-C) and C_{60} test sets. The RMSEs are for predicted effective Hirshfeld volumes that are usually slightly smaller than one (the average of the training set is approximately 0.9). See Fig. 2 d) for a representative range of variation in these data sets.

set there is an increase in error after a certain SOAP cutoff. The descriptors of two different amorphous carbon structures can be very different from one another if a large cutoff is used while the atoms far away from the central atom do not affect the physics of the system significantly. On the other hand, for the C_{60} molecules the descriptors can be quite similar even for large cutoff radii. This is also the reason why the error has not yet reached the minimum for the C_{60} test set and a larger cutoff could be used.

C. Missing contribution to the forces in DFT codes

The effective Hirshfeld volumes are non-variational functionals of the charge density and calculating their gradients thus requires the calculation of the charge density gradients. The DFT code we use as our reference method lacks the implementation of these charge density gradients and the dispersion forces only contain the terms of Eq. (12) that do not include the derivatives of the Hirshfeld volumes. Within the GAP methodology, however, the calculation of the gradients of the Hirshfeld volumes is straightforward and these terms can therefore be included in the model.

In Fig. 4 we demonstrate that our model is capable of reproducing the dispersion forces calculated using the DFT method and that the missing contribution can actually be of the same magnitude as the dispersion forces. Furthermore, we show that the analytical dispersion forces that are produced by our model agree well with the forces that we have calculated using the finite difference (FD) method from the DFT total energies. For these calculations, we used an amorphous carbon structure with 64 carbon atoms in the unit cell and a distorted C_{60} molecule. For the latter, we only included the dispersion interactions within the molecule by making the simulation box large enough and the vdW cutoff radius small enough. For the former, we chose a vdW cutoff radius of 50 Å, which is the default cutoff used by VASP. We have included an analysis of the effect of vdW radius on the dispersion forces in the SI.

For amorphous carbon, the RMSE is approximately 50% larger for the VASP analytical forces than it is for the analytical forces calculated with GAP, when compared to the FD forces that we have calculated with VASP. For the distorted C_{60} structure the errors are an order of magnitude larger for VASP. The database contains a number of different C_{60} structures which enables accurate interpolation with GAP. For the C_{60} molecule the contribution from the Hirshfeld volume gradients to the vdW forces is smaller, compared to a-C, thus making the absolute and relative errors smaller too.

In Fig. 4 c) we show that when we set the Hirshfeld volume gradients to zero, effectively removing the first two terms of Eq. (12), we obtain almost perfect agreement with the analytical dispersion forces given by VASP. GAP is thus able to closely reproduce VASP dispersion forces and, for structures that are well represented in the database, it can indeed predict forces that are *more accurate* than those produced by the reference method. This is one of the strongest results in this paper, and we expect it to generalize to other problems where computation of gradients is hindered by complications in the implementation.

D. Structural transitions in C_{60}

C_{60} has been extensively studied in the literature [41–44], see, e.g., Refs. [45] and [46] for recent overviews. There is fundamental interest, since C_{60} is the molecule with the highest symmetry: 120 symmetry operations altogether (each an element in the icosahedral symmetry group), including 60 rotational symmetry operations. But there is also a strong interest in using C_{60} precursors to synthesize novel forms of carbon by applying different heat and pressure treatments [44, 45].

Individual C_{60} molecules are expected to behave similarly

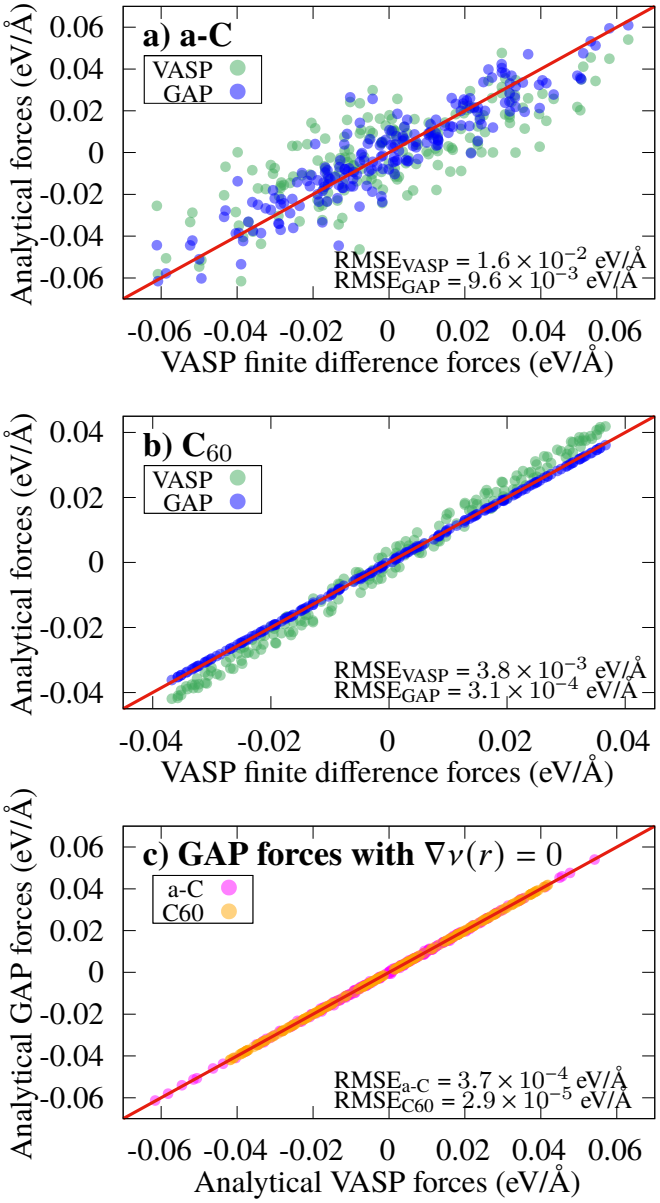

FIG. 4. a) Comparison of dispersion forces (x, y and z components in the same panel) calculated analytically using the TurboGAP implementation and with the FD method from the dispersion energies calculated with VASP. Here the structure is amorphous carbon with 64 atoms in the unit cell and the cutoff radius for vdW interactions is 50 Å. The VASP analytical forces show approximately 50 % larger RMSE than the analytical GAP forces, when both are compared to the finite difference (which we treat as the reference here). b) Comparison of dispersion forces for a slightly distorted C_{60} structure with a vdW cutoff radius of 8 Å to only include atoms in one C_{60} molecule. The RMSE is an order of magnitude larger for VASP when compared to the GAP forces. c) Analytical GAP and VASP dispersion forces for amorphous carbon and C_{60} structures. Here we have switched off the Hirshfeld volume gradient terms to show that we get excellent agreement with VASP dispersion forces, which do not include these gradients in any way.

to soft spheres that interact among each other via vdW forces at long distances. When the C_{60} molecules get close enough that the constituent carbon atoms "see" individual C-C interactions (2b, 3b and mb), rather than an effective spherically symmetric potential centered on the center of mass (CM) of the C_{60} molecule, the situation becomes more complicated. In particular, we are interested in finding out deviations from ideal Lennard-Jones (LJ) fluid behavior at low pressure and in working out the thermodynamic conditions for the transition from a C_{60} fluid to amorphous carbon (a-C) and liquid carbon (l-C) at high pressure and high temperature, respectively. All the simulations in this section use the new C60 GAP and TurboGAP, a cutoff for vdW interactions of 20 Å and $s_R = 0.893$, which was optimized for C as detailed in the SI.

1. *Clustering at low pressure*

At low pressure and low temperature, real systems that behave close to ideal LJ fluids, such as Ar and other noble gases, tend to cluster into closed-packed structures [47, 48]. To elucidate, first, the existence and, subsequently, the structure of C_{60} clusters at low pressure and temperature, we carried out molecular dynamics (MD) simulations with our new C60 GAP potential and the TurboGAP code. We simulated systems made up of 64 C_{60} molecules (a total of 3840 C atoms) in large cubic simulation boxes of size $150 \times 150 \times 150$ Å^3 under periodic boundary conditions (PBC). The positions of the initial configurations were generated by randomizing the position of the CM of each C_{60} molecule while avoiding two C_{60} units from coming too close to one another. We scanned the temperature range from 10 K up to 5000 K, using a Berendsen thermostat [49] with time constant 100 fs and time steps for the MD integration of 2 fs below 500 K and 1 fs at and above 500 K. Each individual MD simulation at a given temperature is run for 500,000 time steps, i.e., 1 ns below 500 K and 500 ps at and above 500 K. Structural analysis is performed after the structural indicators, as described below, have stabilized. This means that we only perform structural analysis on the final portion of the 1 ns or 500 ps trajectories.

To characterize the structure of the clusters we tried using the popular Steinhardt parameter analysis [50, 51], but this proved to be insufficient to clearly resolve the structure of the C_{60} clusters. Therefore, we resorted to a more sophisticated analysis based on mb SOAP descriptors [6, 36], where the same kernel used in Eq. (8) to build GAP potentials is used to characterize the similarity between atomic environments (see Ref. [52] for a recent example in carbon materials). For that purpose, we mapped each C_{60} unit into its CM and computed the radial distribution function (RDF) of the CMs from our MD data. From the location of the first peak in the RDF, we inferred the typical CM-CM first-neighbors distances in the C_{60} fluid, and constructed body-centered cubic (bcc), face-centered cubic (fcc) and hexagonal close-packed (hcp) reference lattices. We then compared the SOAP descriptors of the C_{60} CMs in the simulated clusters to the reference lattices, obtaining numerical scores for each of them. This allows us to establish whether a given C_{60} cluster resembles a bcc, fcc or

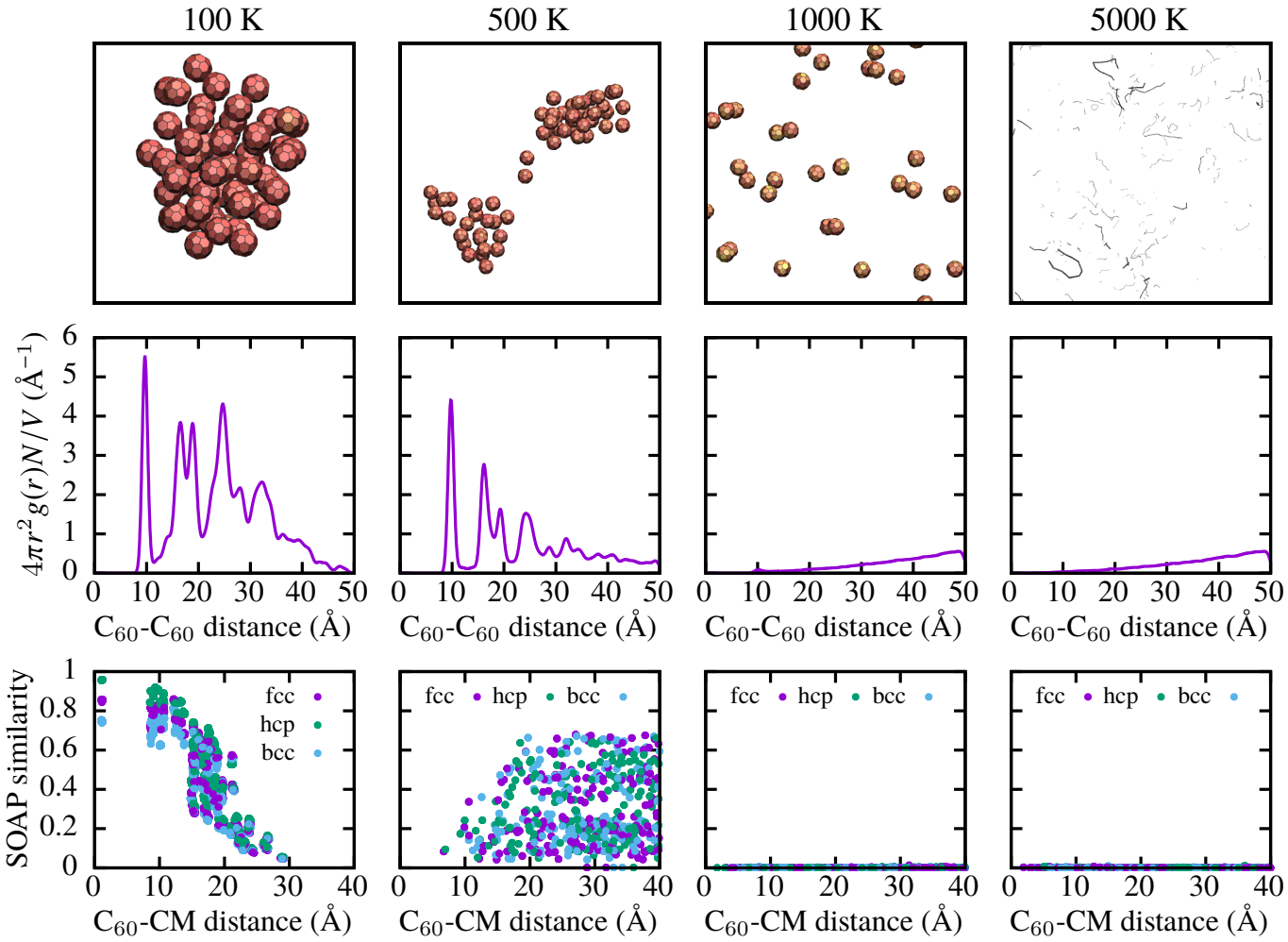

FIG. 5. Top row: different structures resulting from the MD simulations; at low temperature, the C_{60} units coalesce to form clusters, whereas as the temperature is raised to around and above 500 K these clusters become unstable and there is a transition to a C_{60} gas first, and then to an atomic C gas. Middle row: radial distribution function (RDF) $g(r)$ for the center of mass (CM) representation of the C_{60} molecules (i.e., each C_{60} unit is mapped onto its CM and the CMs are used to compute the RDF). The RDF is shown, for convenience, multiplied by the factor required for the integral of the curve from 0 to R to yield the number of C_{60} units at R. Bottom row: SOAP similarity measures between the CM representation and reference simple crystal lattices, plotted as a function of distance between the CM of the C_{60} units and the overall CM of the system. Here the kernel in Eq. (8) is raised to the power of 4 to improve visualization of the difference among crystal structures. Note that the SOAP descriptors used in this analysis are computed for the CM representation, not the full atomic representation.

hcp structure the most. All the results for these low-pressure simulations are summarized in Fig. 5.

If Fig. 5, the top row shows the structures resulting after equilibration for selected temperatures: 100, 500, 1000 and 5000 K. We additionally performed simulations at 10, 20, 50, 200 and 2000 K (not shown). However, the most representative situations are those depicted in the figure. In particular, below 500 K we observed that the individual C_{60} units spontaneously coalesce to form compact clusters. The characteristic distance between C_{60} units is about 9.7 Å, and the structures are clearly ordered, as evidenced from the RDF plot in the middle row (note the RDF is computed for the C_{60} CMs, not the individual C atoms). The bottom row shows the similarity analysis between the actual cluster and the reference fcc, hcp and bcc lattices. The similarity measures, computed for each and all C_{60} units in the system, are plotted as a function of the distance between the CM of the entire system and the CM of the corresponding C_{60} unit. This is necessary since as one moves towards the surface of the cluster the resemblance to an extended crystal is necessarily lost. From this SOAP analysis, we conclude that the structure of the clusters resembles a hcp lattice arrangement, and more so than fcc or bcc. We should note, however, that at these low pressures and temperatures the error in the TS vdW correction due to the missing many-body effects could become significant. The importance of many-body effects on C_{60} clustering should be assessed when a computationally affordable implementation becomes available. Our group is currently working on extending the present model to account for these missing terms.

At around 500 K there is a transition between tightly and loosely connected clusters, with the single cluster that forms at lower energies now splitting into two subunits of similar size.

The SOAP similarity analysis does not provide anymore a clear preference for a specific crystal-like structure, although the RDF still shows a clearly orderly arrangement. Above 500 K the structure transforms into a gas and the vdW interactions are no longer able to keep the C_{60} molecules bound together. The individual C_{60} units remained chemically stable at 1000 and 2000 K (not shown), but they broke apart into their atomic constituents at 5000 K. At this temperature the structure turns into loosely connected open C chains. At this point the system does not retain any memory of having originally been made of C_{60} units.

2. *Solid phase at ambient conditions*

At near ambient conditions, solid C_{60} is experimentally known to adopt an fcc structure [41, 43]. There is also an experimental observation of a rotationally oriented phase, where the C_{60} units are able to librate about equilibrium positions and orientations but not freely rotate, to an orientationally disordered structure where the C_{60} can rotationally diffuse [43]. The temperate of this phase transition has been established experimentally at around 260 K. In this section we try, with mixed success, to reproduce these experimental results, which are a very stringent test of the quality of our model due to the small magnitudes of the energies involved.

All the results of this section are summarized in Fig. 6. First, we try to elucidate the crystal structure predicted by the C60 GAP for C_{60} at ambient conditions. Fig. 6 a) shows a series of successive heating and cooling cycles that we applied to 32-molecule samples arranged in fcc and hcp lattices, coupled to a barostat at 1 bar and a variable thermostat that changes the temperature from 1000 K down to 10 K and back up in 200 ps cycles. The figure shows that fcc is consistently lower in energy than hcp at these conditions. For the low temperature structure (at 10 K), our results indicate a cubic lattice constant for fcc of 13.7 Å, and a cohesive energy of the order of only 1.67 meV/atom (100 meV/molecule) lower (more stable) than hcp. Note that we fixed the c/a lattice parameter ratio of hcp to the ideal value during the simulation; it is therefore possible that the energy difference might decrease even further if allowing for non-ideal c/a ratios (i.e., anisotropic barostating). These tiny energy differences, which are one order of magnitude lower than typical available kinetic energies per degree of freedom at room temperature, explain why under different conditions hcp may become more stable than fcc. We found this to be the case for clusters in the previous section and will also observe this behavior at high pressure and temperature in the next section.

Another intriguing aspect of the structure of C_{60} at ambient conditions is the orientationally ordered to orientationally disordered phase transition [41], which has been thoroughly characterized in Ref. [43]. Namely, below the transition temperature, experimentally observed at around 260 K, the four molecules at the fcc lattice sites are rotationally "locked" in place. Above the transition temperature, they are free to rotate about the fcc lattice sites. The specific orientation at low temperature can be retrieved by a concerted rotation of all four

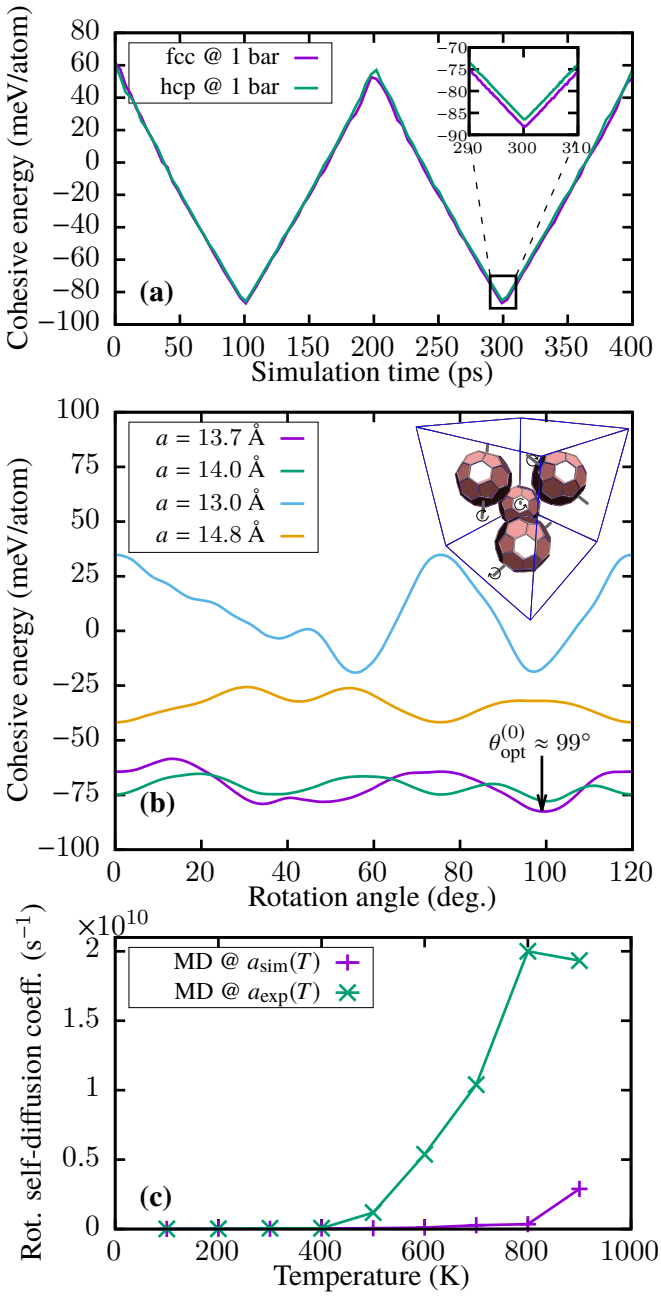

FIG. 6. a) Cohesive energy (referenced to that of an isolated C_{60}) for 32-molecule systems in fcc and hcp structures as the temperature is cycled from 1000 K to 10 K and back to 1000 K in 200 ps cycles, at $P = 1$ bar. b) Potential energy surface of a 4-molecule fcc unit cell as a function of the concerted rotation of angle θ; see text for details. c) Rotational self-diffusion coefficient computed for the 32-molecule fcc system at different temperatures and 1 bar.

molecules, by the same angle θ, around four different local axes (going though each molecule's center of mass). These axes are indicated in the inset of Fig. 6 b), and in more detail in Refs. [43, 46]. In that figure we show the potential energy surface as a function of θ. We find, for the C60 GAP optimal lattice parameter, $a = 13.7$ Å, that there is indeed a local minimum at $\theta = 99°$, very close to the experimental value of 98°.

We also verified with stochastic sampling (1 million random orientations of the four molecules around the fcc lattice sites), that this is also probably the global minimum. We note, however, that this analysis is based on a rigid molecule picture. Indeed, Fig. 6 a) shows that a lower energy configuration can be achieved if the molecules are allowed to deform elastically (as is the case in the course of MD). We also note the very strong dependence of the energy profiles in Fig. 6 b) on the lattice parameter. At the low-temperature experimental value, $a \approx 14$ Å, this minimum becomes less stable compared to other local minima.

Finally, we tried to elucidate the rotational ordered-to-disordered phase transition by calculating the rotational self-diffusion coefficient $D_{\rm rot}$ for 32-molecule fcc systems as a function of temperature, within 20 ps MD trajectories at each T. Within the Stokes-Einstein picture, $D_{\rm rot}$ is expected to grow linearly in T for freely rotating spherical molecules. As expected for C_{60} at low temperature, we find that this is not the case in Fig. 6 c). There is clearly a transition between a rotationally locked phase at low temperature to a rotationally diffuse phase at high temperature, with exponential growth of $D_{\rm rot}$, which increases by several orders of magnitude, in between. This growth of $D_{\rm rot}$ over a range of temperatures is qualitatively in agreement with the existence of a range of temperatures over which C_{60} transitions from fully orientationally ordered to fully orientationally disordered. Within that range, from $\sim$ 90 K to 260 K, certain fractions of molecules attain one of two possible specific relative orientations (so-called "P orientation" and "H orientation"), the precise fraction depending on T and P [46]. Again, we find in Fig. 6 c) that there is a strong dependence of the specific characteristics on the lattice parameter. When we adjust the simulated lattice parameter to the experimental one (which involves an increase of about 2.7% in the lattice parameter), the transition temperature goes down by about 300 K, from $\sim$ 800 K to $\sim$ 500 K. Still, this temperature is far away, quantitatively, from the experimental observation of the phase transition. This may be indicative, together with the underestimation of the lattice parameter, that our approach may be overestimating the vdW interaction. We remain, nonetheless, very satisfied with the overall performance of the present approach, which qualitatively reproduces much of the fine detail of the structure of C_{60}, especially given the tiny energy differences involved.

3. *High pressure/high temperature phase transitions*

We have further studied the structure of the C_{60} fluid under high pressure conditions. In particular, we are interested in demarcating the phase boundary between a molecular fluid/solid made up of individually stable C_{60} units, on the one hand, and a-C and l-C, on the other. In these simulations, we scanned the temperature range where we expect to find the phase boundary, and ran MD simulations at 1000, 2000, 3000, 4000 and 5000 K. We scanned the range of pressures throughout five orders of magnitude on a logarithmic scale, from 1 kbar to 10 Mbar. The simulations were carried out in the following way. 1) A random arrangement of 256 C_{60} units was generated as in the previous section. 2) For each pressure value a 200 ps equilibration was carried out at 1000 K using the Berendsen thermostat with the same parameters as in the previous section, and the Berendsen barostat with time constant of 1 ps and bulk modulus of $\gamma_p = (\sqrt{P/1000}/4.5 \times 10^{-5})$ bar, where P is in units of bar; in the TurboGAP implementation, the bulk modulus is specified in units of the approximate inverse compressibility of water, $1/(4.5 \times 10^{-5}$ bar$^{-1})$. We find that a variable bulk modulus is necessary to accommodate the changing elastic response of the system as it becomes compressed. 3) After the equilibration phase, the barostat is turned off and a box scaling transformation is used to very slowly change the lattice vectors, starting from the last snapshot of the equilibration and ending at the average lattice vectors during the part of the equilibration period where the potential energy has settled. This is done over 20 ps of MD. 4) A fixed-volume simulation is run for another 20 ps with only the thermostat turned on, and structural data (atomic positions) is gathered for further analysis. 5) The last snapshot from step 4) is taken as starting configuration for a new step 2), increasing the temperature by 1000 K. Therefore, for each pressure there is a continuous transformation in temperature from 1000 K to 5000 K, at 1000 K steps, throughout all the stages detailed above.

The results of our simulations are summarized in Fig. 7 and Table II. We find that the region of metastability for a C_{60} condensed phase lies in the lower left quadrant of the metastable phase diagram at and below approximately 3000 K and 100 kbar. At higher temperatures the C_{60} system transforms into either liquid C (l-C) or graphite/graphitic C, depending on the pressure, whereas at higher pressures the transition is to extremely highly sp^3-rich amorphous C (a-C), which we have denoted on the graph as amorphous diamond (a-D). At $T = 5000$ K and 1 Mbar we start to observe crystalline diamond nucleation and grain formation, identified at "polycrystalline" diamond on the figure. For very high pressures (10 Mbar), the density (Table II) rises to more than twice the density of crystalline diamond at ambient pressure ($\sim$ 3.5 g/cm^3).

TABLE II. Selection of structural indicators for the C_{60} system at high pressure: i) Approximate location of the first RDF peak in the CM representation of the C_{60} condensed phase (solid and liquid, i.e., excluding clusters at low pressure) at different temperatures and pressures. ii) Mass densities for the different structures depicted in Fig. 7.

	RDF first peak				
	1000 K	2000 K	3000 K	4000 K	5000 K
1 kbar	9.85 Å	10.13 Å	10.45 Å	N/A	N/A
10 kbar	9.71 Å	10.03 Å	10.12 Å	N/A	N/A
100 kbar	8.86 Å	8.96 Å	N/A	N/A	N/A
$\geq$ 1 Mbar	N/A	N/A	N/A	N/A	N/A
	Mass density (g/cm^3)				
	1000 K	2000 K	3000 K	4000 K	5000 K
1 kbar	1.57	1.27	0.91	0.42	0.33
10 kbar	1.68	1.61	1.47	1.77	1.23
100 kbar	2.09	2.09	2.41	2.40	2.41
1 Mbar	3.89	3.90	3.90	3.90	3.96
10 Mbar	7.23	7.21	7.16	7.17	7.16

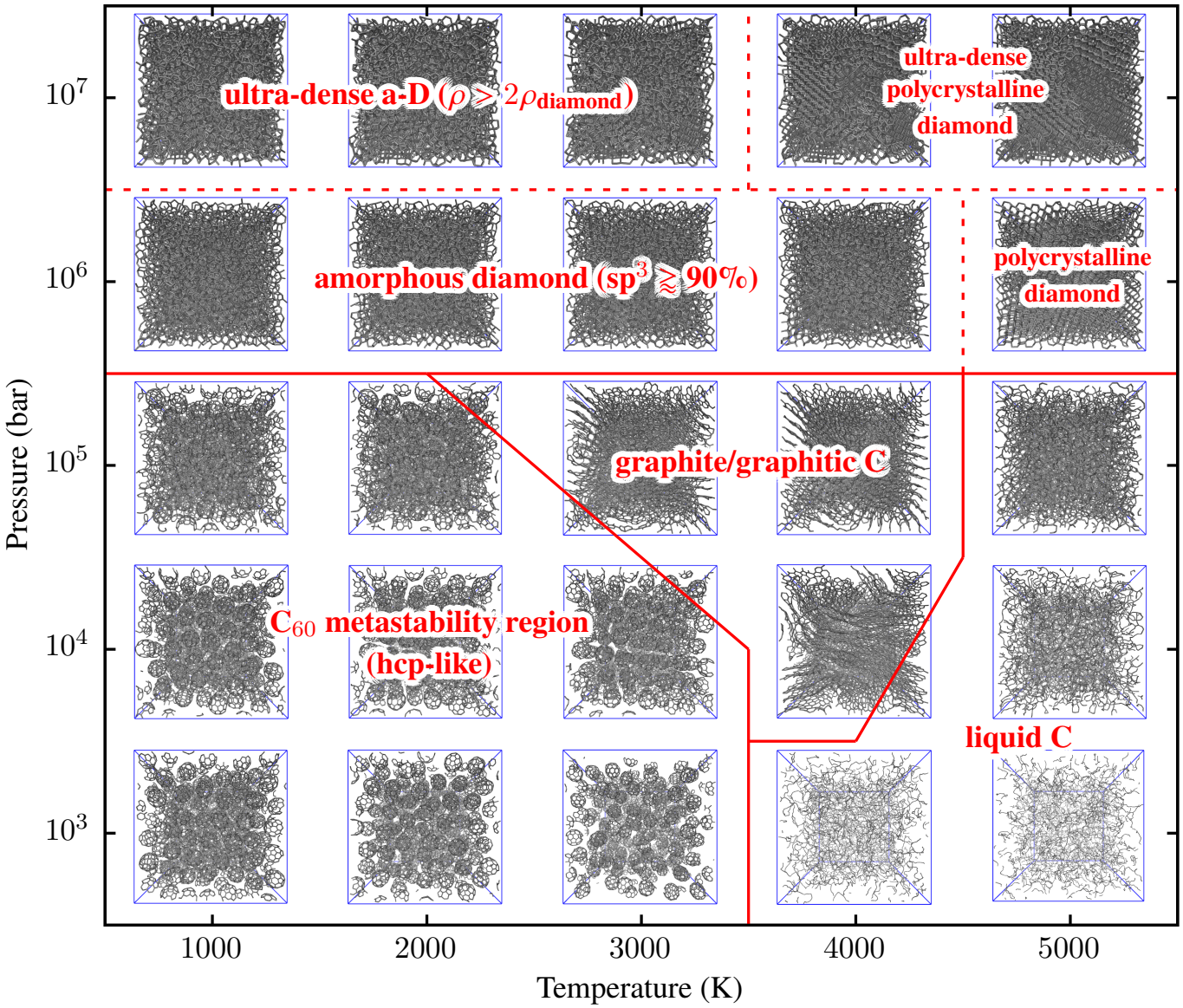

FIG. 7. Metastable phase diagram for the C_{60} system at high pressure (at and above 1 kbar). Transitions to other allotropes of C from the C_{60} seed can be observed as temperature, pressure, or both are increased. See text for a detailed discussion. The atomic structures were drawn with VMD [53, 54] and structure manipulation and analysis were carried out with ASE [30] and different in-house codes.

This high-pressure phase shows, besides high sp^3 content in the vicinity of 75 %, also sizable exotic coordinations where C atoms can be surrounded by 5 and more nearest neighbors. We have thoroughly studied 5-fold coordinated carbon complexes in the context of GAP simulation in our previous work [52]. We have based the neighbor counts on a cutoff scheme, which is commonplace in the literature and usually chosen at 1.85 Å, the first minimum in the RDF at ambient conditions. Due to the extreme compression of a-D in Fig. 7, we adjusted this cutoff to account for the increase in mass density compared to crystalline diamond at ambient conditions, $r_{\mathrm{cut}} = 1.85\ \text{Å} \times \sqrt[3]{\rho_{\mathrm{diamond}}/\rho}$.

The structure of the C_{60} condensed phase was probed using the SOAP similarity metric as in the previous section. We found that the structure is not fcc anymore at the probed pressure/temperature combinations, even though we chose to simulate 256 C_{60} units precisely to accommodate a possible fcc arrangement. While the high-pressure C_{60} structure resembles hcp more than either fcc or bcc, its overall resemblance to hcp decreases, compared to clusters at low pressure. For this reason we have denoted the structure as hcp "like" in Fig. 7.

To elucidate the true nature of the liquid structures, in Fig. 8 we show vibrational density of states (DoS) calculations carried out with the DoSPT code [55–57]. The vibrational DoS spectra show characteristic profiles with strong self-diffusion of the individual C atoms, which is proportional to the zero-frequency DoS [58, 59]. The self-diffusion constant decreases as pressure increases. Together with the high temperature structures identified as "liquid C" in Fig. 7, we have also in-

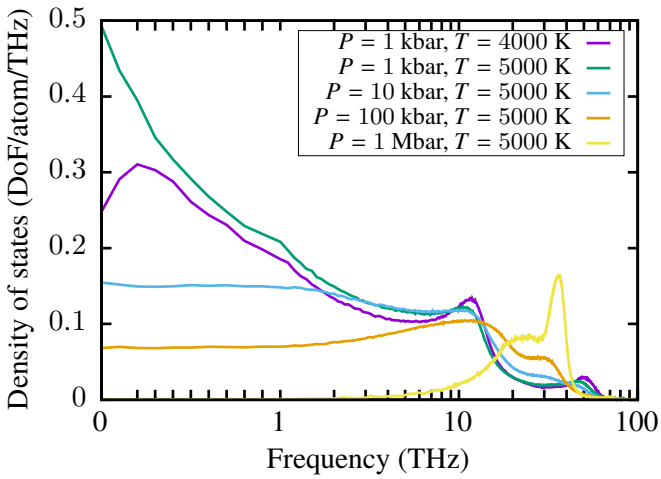

FIG. 8. Vibrational density of states of all the liquid structures identified in Fig. 7 plus the one identified as "polycrystalline diamond", which is used as a reference for a solid.

cluded what we identify as "polycrystalline diamond". The vibrational DoS for that structure shows the bimodal features typical of diamond, and no self diffusion, highlighting that even at 5000 K 10 Mbar of pressure is enough to suppress liquid behavior from the sample. Finally, we note the high frequency peaks at low pressure, at around 50 THz, which correspond to the *sp* chains present in those liquid samples (and which are visible in Fig. 7).

IV. CODE AND POTENTIAL AVAILABILITY

The GAP and TurboGAP codes are freely available online for non-commercial academic research [27, 28]. The GAP_turbo library with modified SOAP descriptors [26] is also freely available under the same terms. QUIP [27] and the C60 GAP [40] are freely available online, for non-commercial as well as commercial research.

V. SUMMARY AND OUTLOOK

We have presented a complete methodology for accurate incorporation of vdW interactions to ML interatomic potentials, based on the Tkatchenko-Scheffler approach. The new approach has been implemented in the GAP and TurboGAP codes and is freely available for academic research. Excellent computational performance can be achieved, with only minor to moderate increase in CPU cost compared to simulations without vdW corrections. We have trained a new ML GAP potential for carbon enabled with the new methodology and optimized it specifically for simulation of C_{60}. The new force field has been validated as an excellent general-purpose carbon potential and utilized to chart the phase transformations taking place in C_{60} under a wide range of temperatures and pressures, reproducing many of the features observed experimentally.

We expect that the new C60 GAP potential will open the door for accurate simulation of carbon materials and C_{60} in particular. There is interest in C_{60} as precursor material for other carbon-based allotropes synthesized under extreme conditions, and we believe that the C60 GAP may guide future experimental efforts in this regard. Methodology wise, the next steps to build up and improve the inclusion of vdW interactions from the current work will focus on reproducing more accurate schemes like PBE+MBD [15] and vdW-DF [60], as well as fitting new vdW-inclusive GAPs for materials other than carbon. Extending our methodology to other non-ionic materials is straightforward, and preliminary work on vdW-inclusive general-purpose GAPs for systems containing C, H and O atoms indicate that the approach works similarly well. On the other hand, systems where significant charge transfer occurs (e.g., when ions are present) will require further development to take into account the nonlocality of the effective Hirshfeld volumes. If they are to remain computationally tractable, and thus amenable to large-scale MD simulation, these new approaches may require a combination of short-range many-body descriptors and simple (e.g., pairwise) long-range ones. All in all, we expect rapid advances in the development of accurate and computationally efficient dispersion correction schemes for ML-driven atomistic modeling in the near future.

ACKNOWLEDGMENTS

The authors acknowledge funding from the Academy of Finland, under projects 310574, 330488 and 329483 (M. A. C.), 321713 (M. A. C, P. H.-L. and H. M.), 308647 (X. C.), 314298 (X. C. and H. M.), and the QTF Center of Excellence program grant no. 312298 (T.A-N.). M. A. C., P. H.-L. and H. M. also acknowledge a Seed Funding grant from the Aalto University Materials Platform. Computing time from CSC – IT Center for Science for the MaCaNa project and from Aalto University's Science IT project is gratefully acknowledged.

[1] R. M. Martin, Electronic Structure (Cambridge University Press, 2004).
[2] A. K. Geim and I. V. Grigorieva, "Van der Waals heterostructures," Nature **499**, 419 (2013).
[3] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, and M. Ceriotti, "Machine learning unifies the modeling of materials and molecules," Sci. Adv. **3**, e1701816 (2017).
[4] V. L. Deringer, M. A. Caro, and G. Csányi, "Machine learning interatomic potentials as emerging tools for materials science," Adv. Mater. , 1902765 (2019).
[5] K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, and K.-R. Müller, eds., Machine Learning Meets Quantum Physics (Springer, 2020).
[6] A. P. Bartók, R. Kondor, and G. Csányi, "On representing

chemical environments," Phys. Rev. B **87**, 184115 (2013).
[7] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, and A. S. Foster, "DScribe: Library of descriptors for machine learning in materials science," arXiv:1904.08875 (2019).
[8] J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[9] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).
[10] S. Grimme, "Semiempirical GGA-type density functional constructed with a long-range dispersion correction," J. Comput. Chem. **27**, 1787 (2006).
[11] A. Tkatchenko and M. Scheffler, "Accurate molecular van der Waals interactions from ground-state electron density and free-atom reference data," Phys. Rev. Lett. **102**, 073005 (2009).
[12] T. Bereau, R. A. DiStasio Jr, A. Tkatchenko, and O. A. Von Lilienfeld, "Non-covalent interactions across organic and biological subsets of chemical space: Physics-based potentials parametrized from machine learning," J. Chem. Phys. **148**, 241706 (2018).
[13] P. D. Mezei and O. A. von Lilienfeld, "Noncovalent quantum machine learning corrections to density functionals," J. Chem. Theory Comput. **16**, 2647 (2020).
[14] T. Morawietz, A. Singraber, C. Dellago, and J. Behler, "How van der Waals interactions determine the unique properties of water," P. Natl. Acad. Sci. USA **113**, 8368 (2016).
[15] A. Tkatchenko, R. A. DiStasio Jr, R. Car, and M. Scheffler, "Accurate and efficient method for many-body van der Waals interactions," Phys. Rev. Lett. **108**, 236402 (2012).
[16] J. Hermann, R. A. DiStasio Jr, and A. Tkatchenko, "First-principles models for van der Waals interactions in molecules and materials: Concepts, theory, and applications," Chem. Rev. **117**, 4714 (2017).
[17] M. Stöhr, T. Van Voorhis, and A. Tkatchenko, "Theory and practice of modeling van der Waals interactions in electronic-structure calculations," Chem. Soc. Rev. **48**, 4118 (2019).
[18] J. P. Perdew, K. Burke, and M. Ernzerhof, "Generalized gradient approximation made simple," Phys. Rev. Lett. **77**, 3865 (1996).
[19] X. Chu and A. Dalgarno, "Linear response time-dependent density functional theory for van der Waals coefficients," J. Chem. Phys. **121**, 4083 (2004).
[20] N. Marom, A. Tkatchenko, M. Rossi, V. V. Gobre, O. Hod, M. Scheffler, and L. Kronik, "Dispersion interactions with density-functional theory: benchmarking semiempirical and interatomic pairwise corrected density functionals," J. Chem. Theory Comput. **7**, 3944 (2011).
[21] M. A. Caro, "Parametrization of the Tkatchenko-Scheffler dispersion correction scheme for popular exchange-correlation density functionals," arXiv:1704.00761 (2017).
[22] A. Otero-de-la Roza, L. M. LeBlanc, and E. R. Johnson, "What is "many-body" dispersion and should I worry about it?" Phys. Chem. Chem. Phys. **22**, 8266 (2020).
[23] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-pu," J. Chem. Phys. **132**, 154104 (2010).
[24] A. P. Bartók and G. Csányi, "Gaussian approximation potentials: A brief tutorial introduction," International Journal of Quantum Chemistry **115**, 1051–1057 (2015).
[25] V. L. Deringer and G. Csányi, "Machine learning based interatomic potential for amorphous carbon," Phys. Rev. B **95**, 094203 (2017).
[26] M. A. Caro, "Optimizing many-body atomic descriptors for enhanced computational performance of machine learning based interatomic potentials," Phys. Rev. B **100**, 024112 (2019).
[27] (), https://libatoms.github.io.
[28] M. A. Caro et al., TurboGAP website and online documentation, [http://turbogap.fi].
[29] E. Snelson and Z. Ghahramani, "Sparse Gaussian processes using pseudo-inputs," in Adv. Neur. In., Vol. 18, edited by Y. Weiss, B. Schölkopf, and J. Platt (MIT Press, 2006).
[30] A. Larsen, J. Mortensen, J. Blomqvist, I. Castelli, R. Christensen, M. Dulak, J. Friis, M. Groves, B. Hammer, C. Hargus, E. Hermes, P. Jennings, P. Jensen, J. Kermode, J. Kitchin, E. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and K. W. Jacobsen, "The Atomic Simulation Environment – A Python library for working with atoms," J. Phys.: Condens. Matter **29**, 273002 (2017).
[31] G. Kresse and J. Furthmüller, "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169–11186 (1996).
[32] G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59**, 1758–1775 (1999).
[33] T. Bučko, S. Lebègue, J. Hafner, and J. G. Ángyán, "Tkatchenko-Scheffler van der Waals correction method with and without self-consistent screening applied to solids," Phys. Rev. B **87**, 064110 (2013).
[34] P. Hernández-León and M. A. Caro, "Cluster-based multidimensional scaling embedding tool for data visualization," **in preparation** ().
[35] P. Hernández-León and M. A. Caro, (), cluster-based multidimensional scaling, [http://github.com/mcaroba/cl-MDS].
[36] S. De, A. P. Bartók, G. Csányi, and M. Ceriotti, "Comparing molecules and solids across structural and alchemical space," Phys. Chem. Chem. Phys. **18**, 13754 (2016).
[37] M. A. Caro, A. Aarva, V. L. Deringer, G. Csányi, and T. Laurila, "Reactivity of amorphous carbon surfaces: rationalizing the role of structural motifs in functionalization using machine learning," Chem. Mater. **30**, 7446 (2018).
[38] V. L. Deringer, M. A. Caro, and G. Csányi, "A general-purpose machine-learning force field for bulk and nanostructured phosphorus," Nat. Commun. **11**, 1 (2020).
[39] B. Cheng, R.-R. Griffiths, S. Wengert, C. Kunkel, T. Stenczel, B. Zhu, V. L. Deringer, N. Bernstein, J. T. Margraf, K. Reuter, and G. Csányi, "Mapping materials and molecules," Accounts Chem. Res. **53**, 1981 (2020).
[40] H. Muhli and M. A. Caro, "GAP interatomic potential for C_{60}," Zenodo (2021), DOI: 10.5281/zenodo.4616343.
[41] Y. Quo, N. Karasawa, and W. A. Goddard, "Prediction of fullerene packing in C_{60} and C_{70} crystals," Nature **351**, 464 (1991).
[42] R. D. Johnson, C. S. Yannoni, H. C. Dorn, J. R. Salem, and D. S. Bethune, "C_{60} rotation in the solid state: Dynamics of a faceted spherical top," Science **255**, 1235 (1992).
[43] J. D. Axe, S. C. Moss, and D. A. Neumann, "Structure and dynamics of crystalline C_{60}," Solid State Phys. **48**, 149 (1994).
[44] B. Sundqvist, "Fullerenes under high pressures," Adv. Phys. **48**, 1 (1999).
[45] C. Pei and L. Wang, "Recent progress on high-pressure and high-temperature studies of fullerenes and related materials," Matter Radiat. Extremes **4**, 028201 (2019).
[46] B. Sundqvist, "Carbon under pressure," Phys. Rep. (2021), https://doi.org/10.1016/j.physrep.2020.12.007.

[47] A. Rytkönen, S. Valkealahti, and M. Manninen, "Phase diagram of argon clusters," J. Chem. Phys. **108**, 5826 (1998).

[48] P. Schwerdtfeger, N. Gaston, R. P. Krawczyk, R. Tonner, and G. E. Moyano, "Extension of the Lennard-Jones potential: Theoretical investigations into rare-gas clusters and crystal lattices of He, Ne, Ar, and Kr using many-body interaction expansions," Phys. Rev. B **73**, 064112 (2006).

[49] H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. R. H. J. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," J. Chem. Phys. **81**, 3684 (1984).

[50] P. J. Steinhardt, D. R. Nelson, and M. Ronchetti, "Bond-orientational order in liquids and glasses," Phys. Rev. B **28**, 784 (1983).

[51] T.-Q. Yu, P.-Y. Chen, M. Chen, A. Samanta, E. Vanden-Eijnden, and M. Tuckerman, "Order-parameter-aided temperature-accelerated sampling for the exploration of crystal polymorphism and solid-liquid phase transitions," J. Chem. Phys. **140**, 06B603_1 (2014).

[52] M. A. Caro, G. Csányi, T. Laurila, and V. L. Deringer, "Machine learning driven simulated deposition of carbon films: from low-density to diamondlike amorphous carbon," Phys. Rev. B **102**, 174201 (2020).

[53] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," J. Molec. Graphics **14**, 33 (1996).

[54] (), http://www.ks.uiuc.edu/Research/vmd/.

[55] M. A. Caro, T. Laurila, and O. Lopez-Acevedo, "Accurate schemes for calculation of thermodynamic properties of liquid mixtures from molecular dynamics simulations," J. Chem. Phys. **145**, 244504 (2016).

[56] M. A. Caro, O. Lopez-Acevedo, and T. Laurila, "Redox potentials from ab initio molecular dynamics and explicit entropy calculations: application to transition metals in aqueous solution," J. Chem. Theory Comput. **13**, 3432 (2017).

[57] M. A. Caro, DoSPT code, [http://dospt.org].

[58] D. A. McQuarrie, Statistical Mechanics (Harper & Row, New York, 1976).

[59] S.-T. Lin, M. Blanco, and W. A. Goddard III, "The two-phase model for calculating thermodynamic properties of liquids from molecular dynamics: Validation for the phase diagram of Lennard-Jones fluids," J. Chem. Phys. **119**, 11792 (2003).

[60] M. Dion, H. Rydberg, E. Schröder, D. C. Langreth, and B. I. Lundqvist, "Van der Waals density functional for general geometries," Phys. Rev. Lett. **92**, 246401 (2004).